# Estimation of voice source and vocal tract characteristics based on multi-frame analysis

# Estimation of Voice Source and Vocal Tract Characteristics Based on Multi-frame Analysis

**Yoshinori SHIGA and Simon KING**
Centre for Speech Technology Research
University of Edinburgh
yoshi@cstr.ed.ac.uk

## 1. Motivation

### Problem

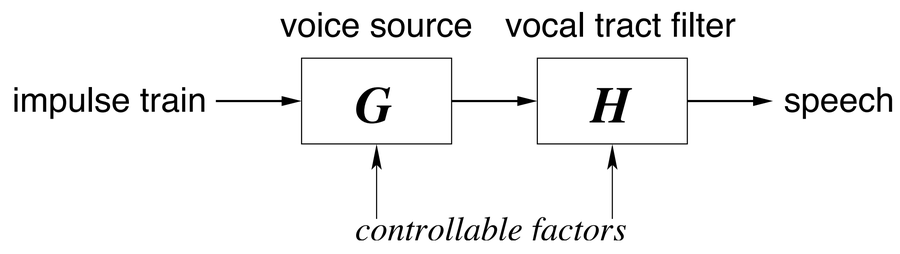
From the viewpoint of filter design in acoustics, it seems **almost impossible** to **simultaneously estimate** the response of a system (*vocal tract*) **and** its input (*voice source*) **only from** its output (*speech*).

### Two types of conventional solutions

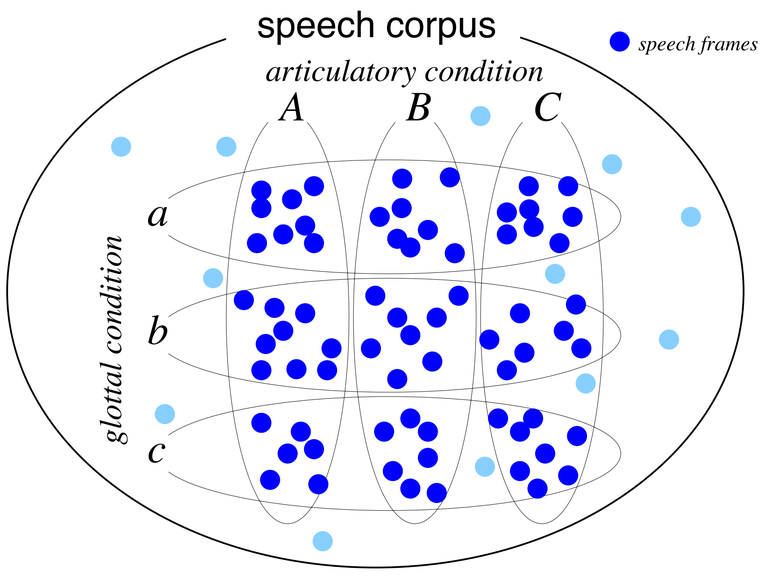
A. The vocal-tract response is estimated under the constraint of **simplistically modelled** voice-source waveform.
B. The voice-source waveform is estimated by filtering speech signal through the inverse of a vocal-tract response approximated by rather **simplistic models**.

## 2. The idea

### Corpus-based simultaneous estimation




Assumed here is **a linear model** composed of two cascade components.




*If the transfer function of each component is controlled by a set of factors which is completely independent of that of the other component,* those two functions can be **approximately separable** by **iterative approximation** using **a large corpus** well-balanced with those factors of both the components.

### Assumptions

A. The source response $\boldsymbol{G}$ changes depending only on the **speech $F_0$ and power**.
B. The filter response $\boldsymbol{H}$ changes depending only on the **vocal tract shape**.

## 3. Multi-frame analysis

### Multi-frame cepstral analysis (MFCA)

**MFCA** is a method to estimate a **spectral envelope** using **multiple frames** of voiced speech by **cepstrally smoothing their harmonics** (not their spectral envelopes).
**(see Poster # for details)**

## 4. Simultaneous estimation of source and filter responses

### Controllable Factors (CFs)

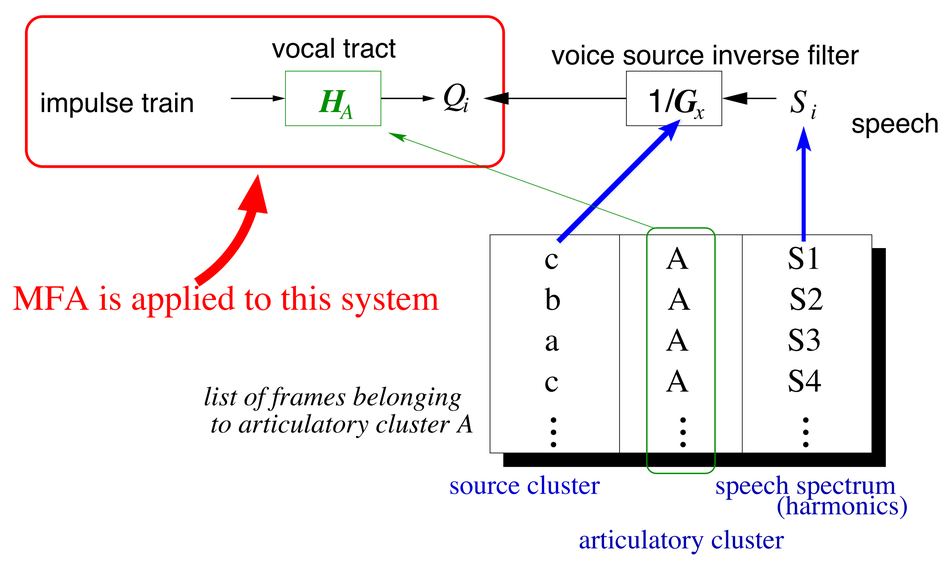
**voice source component**— $F_0$ and power of speech
**vocal tract component**— **electromagnetic articulograph (EMA)** data

### Grouping frames with similar CFs

The following two types of **clusterings** are applied to the same corpus.

- Based on the *EMA data*, all the voiced frames included in the corpus are divided into $K$ clusters (**articulatory clusters**), $C_h^{(i)}(i = 1, 2, 3, ..., K)$, so that each of the clusters consists of frames with similar articulatory settings
- Based on the *$F_0$ and speech power*, all the frames are divided into $L$ clusters (**source clusters**), $C_g^{(j)}(j = 1, 2, 3, ..., L)$, so that each of the clusters consists of frames with similar $F_0$ and power values.

### Estimating vocal tract responses

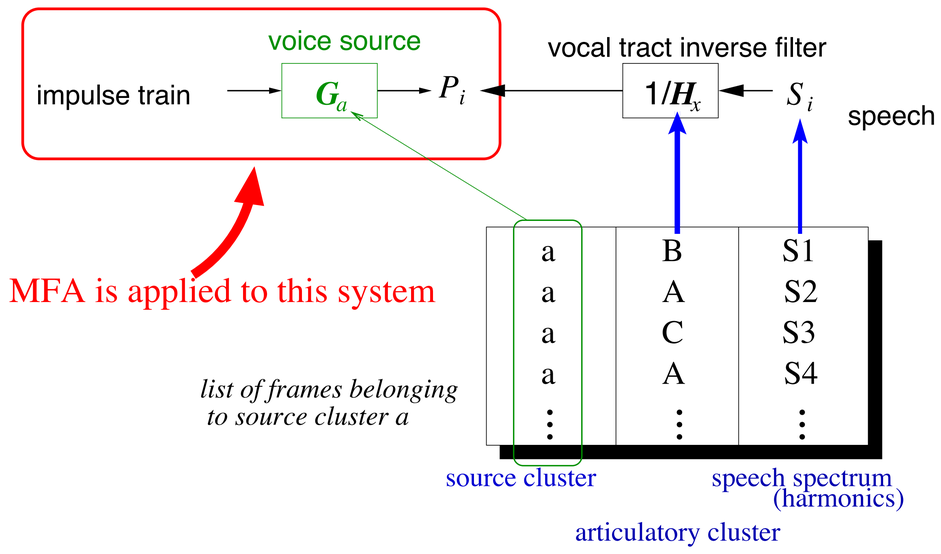



**Step 1:** Log-amplitudes of harmonics calculated from the voice source cepstrum, $\boldsymbol{c}_g^{(j)}(j = 1, 2, 3, ..., L)$, are subtracted from the observed log-amplitude, $\boldsymbol{a}_k$, as follows:

$$\boldsymbol{q}_k = \boldsymbol{a}_k - \boldsymbol{B}_k \boldsymbol{c}_g^{(R_g(k))}, \quad j = R_g(k) \Longleftrightarrow k \in C_g^{(j)}$$

**Step 2:** For each articulatory cluster $C_h^{(i)}(i = 1, 2, 3, ..., K)$, $\boldsymbol{c}_h^{(i)}$ is calculated by applying MFCA to $\{\boldsymbol{q}_k | k \in C_h^{(i)}\}$.

## Estimating voice source responses




**Step 3:** Log-amplitudes of harmonics calculated from the vocal tract cepstrum, $\boldsymbol{c}_h^{(i)}$, are subtracted from the log-amplitudes of the observed harmonics, $\boldsymbol{a}_k$, as follows:

$$\boldsymbol{p}_k = \boldsymbol{a}_k - \boldsymbol{B}_k \boldsymbol{c}_h^{(R_h(k))}, \quad i = R_h(k) \Longleftrightarrow k \in C_h^{(i)}$$

**Step 4:** For each articulatory cluster $C_g^{(j)} (j = 1, 2, 3, ..., L)$, $\boldsymbol{c}_g^{(j)}$ is calculated by applying MFCA to $\{\boldsymbol{p}_k | k \in C_g^{(j)}\}$.

In the above equations,

$$\boldsymbol{B}_k = \begin{bmatrix} 1 & 2\cos(\Omega_k^1) & 2\cos(2\Omega_k^1) & \cdots & 2\cos(p\Omega_k^1) \\ 1 & 2\cos(\Omega_k^2) & 2\cos(2\Omega_k^2) & \cdots & 2\cos(p\Omega_k^2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2\cos(\Omega_k^{N_k}) & 2\cos(2\Omega_k^{N_k}) & \cdots & 2\cos(p\Omega_k^{N_k}) \end{bmatrix}, \quad \Omega_k^l = 2\pi f_k^l T$$

## Iterative estimation

The procedure starts with the initial condition, $\boldsymbol{c}_g^{(j)} = \boldsymbol{0}$ (for all $j$). Estimation error from MFCA is evaluated and the procedure is terminated if the error converges. If not, Step 1-4 is applied repeatedly.

# 5. Experiment

## Experimental condition

- We used a **MOCHA-TIMIT corpus** with parallel acoustic-articulatory information.

***MOCHA-TIMIT corpus***

| | | |
|---|---|---|
| speaker | | female (fsew0) |
| number of sentences | | 460 |
| sampling rate | speech | 16.0 kHz |
| | EMA data | 500 Hz |

- We extracted 87208 voiced frames from the corpus using the following analysis.

***Harmonic estimation***

| | | |
|---|---|---|
| method | | weighted LSM (*Stylianou,2001*) |
| analysis window | type | Hanning |
| | width | 20.0 ms |
| | spacing | 8.0 ms |

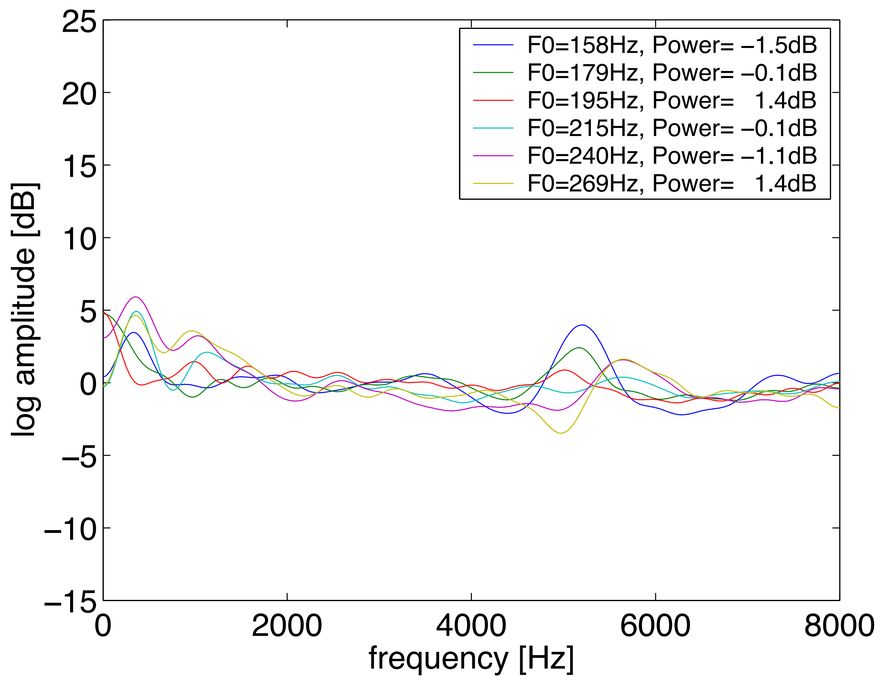
- We applied MFCA under the following conditions.

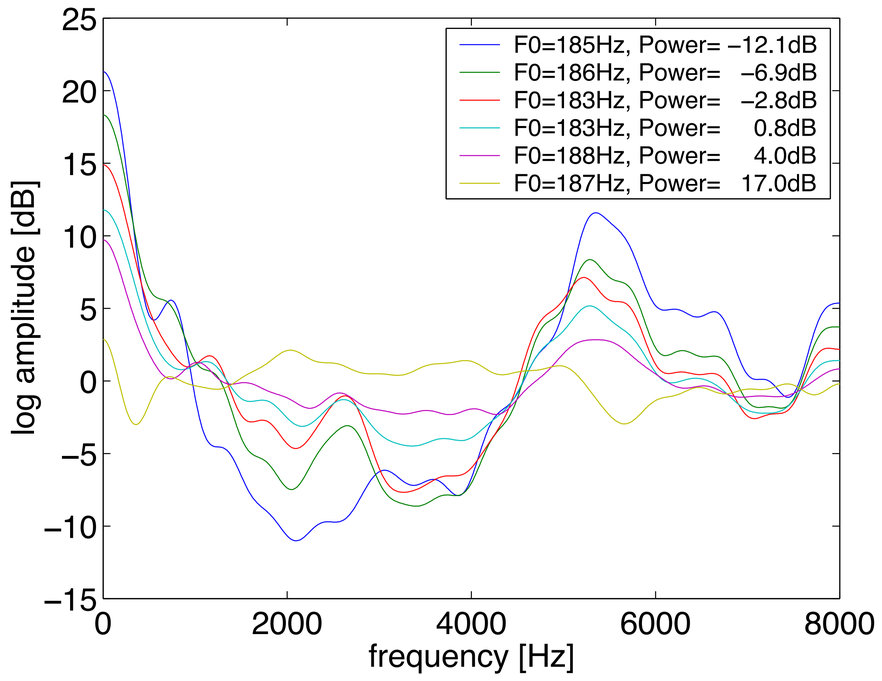
***Multi-frame cepstral analysis***

| | | |
|---|---|---|
| clustering method | | LBG clustering (*Linde et al.,1980*) |
| number of clusters | articulatory cluster | 4096 |
| | source cluster | 64 |
| order of cepstrum | | 32 |

# 6. Results and discussion

## Voice source responses




## Discussion

- The voice-source response does not change significantly with $F_0$, though **does with power when the power is low**.
- The response with low speech power shows **a large amount of energy at the first harmonic** (fundamental) and suppressed higher harmonics in the low frequency band. This tendency is very much in agreement with reports (e.g. Miller,1959) that the glottal waveform becomes **more sinusoidal** when the power of voice is low.
- In the high frequency band, the response with low speech power has much power compared to the amplitude in the low frequency band. We think that it shows the relative increase of noise level since SN ratio reduces when voice power is low.

# 7. Conclusions

## Further discussion

- Strictly, the proposed method does not completely separate the characteristics of voice source and vocal tract filter.
- The method allows us, however, to **control each response independently**, and those responses can be **estimated automatically** from the corpus, which is useful for speech synthesis.
- Informal re-synthesis experiments showed that **intelligible, high quality speech** was generated at least for two speakers. (Based on **sinusoidal synthesis** using cepstra and the original $F_0$ contour)